SYSTEM AND METHOD FOR CLUSTERING A SET OF RECORDS

ABSTRACT OF THE INVENTION

A record clustering system provides a computationally inexpensive method of accurately clustering data records containing structured raw data requiring only two passes over the data. Each data record contains a sequence of attribute values of corresponding attributes. For each attribute, a characteristic value is calculated by evaluating the attribute values of that attribute across the data records. For each attribute value, a deviation from the corresponding characteristic value is calculated. The attributes of each record are sorted based on the deviations to provide a sequence of attributes used as a key for clustering. A user may select criteria for evaluation of the keys for clustering of the data records. The clustering result is refined by searching of best matching keys in other clusters for the records of the smallest cluster. In this manner, the records contained in the smallest cluster are distributed to other clusters, reducing the total number of clusters.